5-2023

Analysis of a Federated Learning Framework for Heterogeneous Medical Image Data: Privacy and Performance Perspective

Julia Brixey
University of Arkansas, Fayetteville

Analysis of a Federated Learning Framework for Heterogeneous Medical Image Data: Privacy and Performance Perspective

An Undergraduate Honors Thesis

in the

Department of Computer Science and Computer Engineering

College of Engineering

University of Arkansas

Fayetteville, AR

April 2023

by

Julia Brixey

This thesis is approved by:

Thesis Advisor:

Signature: _____ Date: 04/27/2023

Printed: Dr. Ukash Nakarmi

Thesis Committee:

Signature: _____ Date: 04/27/2023

Printed: Dr. Susan Gauch

Signature: _____ Date: 04/27/2023

Printed: Dr. Thi Hoang Ngan Le

Abstract

The massive amount of data available in our modern world and the increase of computational efficiency and power have allowed for great advancements in several fields such as computer vision, image processing, and natural languages. At the center of these advancements lies a data-centric learning approach termed deep learning. However, in the medical field, the application of deep learning comes with many challenges. Some of the fundamental challenges are the lack of massive training datasets, unbalanced and heterogenous data between health applications and health centers, security and privacy concerns, and the high cost of wrong inference and prediction. One of the interesting questions of data-centric learning in the medical field is whether we can leverage the heterogenous data available in several medical facilities in a combined way without actually sharing the data between the institutes and preserving the security and privacy of patients. One way to address this question is through the use of the federated deep learning technique. In federated deep learning, the "learning" from each local deep learning model trained on a small, distinct dataset is shared with a global model instead of sharing the actual data and hence does not violate any security and privacy concerns.

In this study, we aim to evaluate the efficiency of the federated learning approach on classification tasks in the medical image domain. Learning in the medical image domain is often more challenging and distinct than that of natural images because of heterogeneity in the data and the unavailability of clear, discernable discriminant features between images of different classes. To this end, we investigate federated learning in medical images in terms of model architecture and data complexity. Through our experiments, we will also investigate the effect that federated learning will have on each local model's performance, and how it affects model generality to external datasets.

Table of Contents

I. Introduction

As computer computational speed, memory capacity, and sophistication have exponentially increased in the past decades, we have seen great strides in the field of deep learning [4]. The modern neural network approach, modeled on human brain processing patterns, has become a staple in the field of artificial intelligence and garnered much attention over recent years [21]. The widespread success and implementation of deep learning applications can be attributed in part to the availability of large datasets. A massive amount of data is generated daily by commercial and private activities [6], and when organized and analyzed properly, fuels deep learning progress. Huge advancements have been made in the realm of computer vision [17]. In recent years, deep learning methods utilizing computer vision have been seen to outperform even state-of-the-art machine learning techniques [23]. Computer vision has been adapted to be integrated into a wide variety of industry sectors and has been implemented in many practical fields, including agriculture, transportation, and retail [1, 18]. Notably, computer vision has also been increasingly utilized in healthcare. Image segmentation, activity recognition, predictive analysis, and classification systems have been applied in many areas, including diagnostic settings, surgery preparation, and therapy development [7]. Advancements have already been made to drastically improve the quality of healthcare provided, lift some of the burdens from overworked medical professionals, and assist in the advancement of important medical research [22]. As this technology grows more specialized and continues to improve, it stands to completely revolutionize the healthcare field.

Producing grand results using deep learning models comes at a cost, however. The explosive success of deep learning is directly tied to the availability of large and high-quality training datasets. Unlike many traditional machine learning or shallow networks that have been found to reach a maximum threshold of accuracy, even when trained on larger datasets [16], deep

learning systems have been shown to have the capacity to continue learning up to a higher accuracy threshold. Without access to this type of data, many systems, including computer vision systems, cannot reach the high levels of accuracy expected from deep learning architectures. Because these architectures are so reliant on large datasets, developing and training a specialized deep-learning algorithm requires a lot of computational power, even when such a dataset is accessible. Deep learning architectures must also have some degree of interpretability in order to be reliably implemented in real-world applications, especially in the medical field [20]. As deep learning typically relies on a black-box approach, it is not always obvious how or why models reach certain conclusions. With low interpretability, it is hard to trust that these models are attending to the features of a dataset that are intuitively important to a human classifier. Even for consistently high-performing models, it is unsure how they would perform if given more diverse datasets or entrusted with real-time tasks.

The large datasets that deep learning models are reliant on are hard to come by for medical-related tasks due to the difficulty in manufacturing large amounts of relevant medical data. Even when data is available, it is a time-consuming and tedious process for medical professionals to provide labels for this data [2]. Also, because medical data labeling is a subjective process, it is a concern that the label that one professional may assign to a data object would not be corroborated by another individual professional. This potential for variability in data labeling contributes even more to the problem of model interpretability. Another roadblock in the development of medical datasets is patient privacy concerns. Medical data is sensitive, and many patients are not comfortable with their data being shared with the large corporations or tech companies responsible for much of the research on artificial intelligence in the medical field [12]. Without the strict rules governing the management of patient data in hospitals in place at

private corporations, the medical data given to these corporations have the potential of being misused or transferred without the patients' knowledge. Even the transfer of medical data itself poses a privacy risk. If a breach were to occur, with current advances in reconstructive technology, even anonymized data could be used to identify specific individuals using advanced reidentification methods [12].

All of these factors lead to a shortage of quality medical image datasets, which poses a problem for healthcare-oriented deep learning research. In order to address these challenges, many researchers have begun to lean towards a more data-centric approach to developing deep learning architectures, as opposed to the traditional model-centric approach. In a data-centric architecture, the dataset is the key component, and a model is developed around the specific needs and features of the data [14]. This approach has become necessary when wanting to develop models to be trained on medical data, as the scarcity of medical datasets forces researchers to prioritize finding a way to utilize the data they do have access to effectively.

Another direction that researchers have taken in order to address the challenge of collecting quality medical data is to utilize a federated learning approach. Federated learning is a practice initially proposed by Google and was specifically designed to address privacy concerns and reduce the risk of data leakage, specifically for the data gathered from mobile devices. In their initial proposal, federated learning is described as a way for isolated models to send updates to a shared global model as they train on their unique local dataset [11]. Data is never shared with the global model, and as it receives updates from all the participating local models, it averages and applies them to its own network. The updated global model is then shared with all local models, and they continue the training and updating cycle as needed. In theory, even though the global model does not have access to any data for training purposes, the updates from the

local models should transfer the knowledge that they have learned from their own datasets, resulting in a more generalizable and sophisticated network. This eliminates the need for transferring and storing data centrally. In a medical setting, this approach would drastically reduce privacy concerns, as patient information is never transferred or shared in any capacity, except with its corresponding local model. Since the introduction of federated learning, many current studies have investigated ways to optimize the practice. Notably, because the datasets of the local models only ever contain local information, there is a tendency for this training data to be unbalanced and non-independent or identically distributed (non-IID) [3]. Traditionally, this type of data would pose an issue in training, as unbalanced and non-IID data often result in biased models. However, with optimization, the implementation of federated learning could update these local networks with generalizable parameters, allowing them to be applicable to more diverse datasets.

In this study, we aim to:

1) Evaluate the efficiency of federated learning on medical imaging datasets for different model complexity and data complexity.
2) Compare the efficiency of the global federated learning-generated model to each of the local models' performance, and
3) Evaluate the generality of the local models' performance to the other local datasets, after being trained with and without the implementation of federated averaging.

To accomplish these goals, we will use two simple but different deep learning models with different complexity, two Magnetic Resonance Imaging (MRI) datasets with different complexity i) 3-dimensional Structural MRI volume images [19] obtained from patients with several mental disorders and ii) 2-dimensional MR head images with different degrees of motion

artifacts, ranging from containing slight motion artifacts but still of diagnosable quality to images with excessive motion artifacts and no diagnosable quality [13]. We use different deep learning models for classification tasks on the different datasets under the federated learning settings. In sections II, III, and IV, we will provide a brief introduction to the federated learning approach, datasets, and deep learning models, respectively used in this study. In section V, we present details of our experiment settings and the results in section VI. Section VII provides the discussions and observations from our experiments and results and finally section VII provides conclusions from this study.

II. Federated Learning Algorithm

There are many ways to implement federated learning on a system of models and data. A popular and successful approach is the implementation of federated stochastic gradient descent (FedSGD). In this approach, a C-fraction of local models referred to as "clients" in the original study, is selected to perform computations, with C = 1 corresponding to a full global batch (all local models are being used). For each client k in C, the mini-batch size of local data used for computations is represented by B. In this case, when B = ∞, it indicates that the entire local dataset is being used for training. The number of training rounds each client k performs on its local dataset before an update is sent to the global model, referred to as the "server" in the original study, is represented by E. In a traditional FedSGD implementation, C = 1, B = ∞, and E = 1. With these parameters set, the average gradient for the server, w_t, is calculated for each client k using the following algorithm, $g_k = \nabla F_k(w_t)$, where $F_k(w_t)$ represents the averaged loss of the k clients on their local parameters w, calculated using algorithm

$F_k(w) = \frac{1}{n_k} \sum_{i \epsilon P_k} f_i(w)$ [3]. In this algorithm client loss, $f(w)$, is found using n_k, the number of data items in each client's local dataset, and P_k, is the set of local data. Once g_k is calculated, the server aggregates the gradients and applies an update using the algorithm $w_{t+1} \leftarrow w_t \leftarrow \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k$ for a set learning rate η [3]. The federated averaging approach is a variation of FedSGD. In federated averaging, each local client can be trained for multiple rounds ($w_{t+1} \leftarrow w_t \leftarrow \eta \nabla F_k(w^k)$) before passing an update to the server (E = 2 or more) [3]. For this study, we will be using a federated averaging algorithm in order to increase local client computation and evaluate its effect on the global outcome.

III. Data

Our first round of experiments utilized the SRPBS Multi-disorder MRI Dataset [19]. This dataset comprises 3D resting-state functional MRI images, 3D T-1-weighted structural MRI images, and fieldmaps for a set of 1627 patients. The data was gathered from 12 different hospitals, referred to as "sites", and is classified by patient diagnosis, with nine diagnoses represented in the dataset. Each MRI image was face-masked (or "defaced") in order to preserve patient privacy and reduce the risk of facial reconstruction in the case of a data breach. We focused on working with the set of T-1 weighted structural MRI images. In our experiments, we created "local" datasets for six sites. For each chosen site, we classified data by patient diagnosis, with healthy controls, major depressive disorder, and schizophrenia being the chosen diagnoses represented in the datasets.

We also utilized data gathered from the Movement-Related ARTefacts (MR-ART) dataset [13]. This dataset contains 3D T1-weighted structural MRI images of 148 healthy patients. The facial features in the MRI images of this dataset were also removed. Three MRI images were collected for each patient. One while they remained still, one with slight motion, and one with more excessive motion. The data is classified by the three different levels of motion. From this 3D MR-ART dataset, we also created a new 2D T1-weighted structural MRI dataset by slicing the 3D MRI volumes into 2D images (referred to as the 2D MR-ART dataset). To create this dataset, we selected 5 individual patients from each class, resulting in 3,843 total images (1282 per class).

We used the popular Modified National Institute of Standards and Technology (MNIST) database to validate the performance of our constructed models and the federated averaging algorithm [5]. The MNIST training dataset consists of 60,000 images of handwritten numbers between 0-9. This data is classified by the number represented.

IV. Model Architecture

The models used in our experiments were chosen in order to best reflect the benefits of federated averaging in wide general use cases. Because we are studying the use of a data-centric approach, we did not design and implement any novel architectures or make drastic changes to any models chosen.

We made use of a ResNet3D-18 in order to perform classifications on our 3D datasets [8, 10]. This 3D neural network is adapted from the popular ResNet (used for 2D images) and has shown considerable success in 3D medical image classification tasks [25]. The ResNet3D-18

model was also adapted for use in classifying our 2D image datasets, resulting in a standard ResNet-18 model [9].

For specific use with our 2D image datasets and to test the adaptability of the federated averaging algorithm to different model types, we implemented two basic classification networks. The first model, called "Net2nn", consisted of three linear and fully connected layers. This model was integrated from a sample implementation of the federated averaging algorithm and has been shown to perform well at simple classification tasks [15]. We also made use of an adapted LeNet with three convolution layers and two fully connected layers. [24]. These models, while simplistic, are known to achieve acceptable levels of classification accuracy, allowing us to shift our focus to the performance of the federated averaging algorithm.

V. Experiments

Our first set of experiments were designed to test the performance of the federated averaging algorithm on more complex medical imaging data. For each local site in the SRPBS Multi-disorder MRI Dataset, we established a baseline of performance using the ResNet3D-18 by training each site with their local dataset for a total of 100 epochs. The data from these sites were then compiled into a global dataset. We then randomly assigned each site, represented by the term "center" in our federated averaging algorithm, a new selection of independent and identically distributed (IID) data, with each class being represented by the same number of samples. We performed the federated averaging experiment with the six local centers by training each center on their local data using ResNet3D-18 for ten epochs before averaging ($C = 1$, $B = \infty$, and $E = 10$), for a total of ten iterations of averaging.

We then repeated the same experiment to test the performance of federated averaging on the 3D MR-ART data. Because this data was not separated by site, we established the baseline performance on the ResNet3D-18 training the model on the entire global dataset for 100 epochs. For use in our federated averaging algorithm, we separated this global dataset into six local datasets for six established centers to train on. As in the previous experiment, the data was independent and identically distributed (IID) and represented an equal number from each class. We ran the federated averaging algorithm using the same parameters of the previous experiment ($C = 1$, $B = \infty$, and $E = 10$, on ResNet3D-18), for a total of ten iterations of averaging.

Our next set of experiments involved the use of the 2D MR-ART dataset. A baseline of performance for the entire global dataset was established through training the ResNet-18 model for 100 epochs. We then split the data into evenly distributed local datasets, with the data independently and identically distributed (IID) to six centers. We then proceeded to run the same federated averaging experiment as described above (using parameters $C = 1$, $B = \infty$, and $E = 10$, for 10 iterations of averaging) with training being performed on the ResNet-18 model.

In our next round of experiments, we utilized the six evenly distributed local datasets generated from the 2D MR-ART dataset for use by the centers of the federated averaging algorithm. We repeated the federated averaging experiment with the same center data distribution and federated averaging process using the Net2nn model and the adapted LeNet model (using parameters $C = 1$, $B = \infty$, and $E = 10$, for 10 iterations of averaging). We saved the models trained for each center from the federated averaging experiment using the adapted LeNet model and tested the validation accuracy of each model on the other centers' data. We then trained the LeNet model on each center's local dataset for 100 epochs without implementing federated averaging. We tested the validation accuracy of each center on the other five center datasets for

these saved models as well. These experiments were performed to not only test the performance of the federated averaging-generated server but also to evaluate the effect that the federated averaging algorithm has on each center's performance. We aim to establish if the process of federated averaging makes the center models more generalizable to a more diverse set of data.

We performed the same set of federated averaging experiments as defined above on the MNIST data. As in all other experiments, the MNIST data was independently and identically distributed (IID) to each of the six centers, with the same numbers of each class being represented. The same federated averaging experiment was then performed (C = 1, B = ∞, and E = 10, for 10 iterations of averaging) using the Net2nn and adapted LeNet models.

VI. Results

In our first round of experiments, we found that the ResNet3D-18 was unable to identify any distinguishable characteristics in the data taken from the SRPBS Multi-disorder MRI Dataset. With each training run, we determined that the model was guessing the same class each time, no matter what the input data was. Even when the data was evenly split by class for use in the federated averaging algorithm, the ResNet3D-18 was not computationally powerful enough to establish any feature distinction. The resulting trained server model and individual center models generated from the federated averaging algorithm were shown to also guess the same class each time. Because the goal of this study is to evaluate the effects of federated averaging, and not to build a complicated model capable of performing complex classifications, we were forced to discard this dataset in order to advance with our experiments. To this end, we conclude that either the features presented in the qualitative T1, and T2 weighted MRI images do not have

discernable discriminant features that separate one mental disorder image from another, or the individual local models need to be more complex and equipped with more learnable parameters.

The ResNet3D-18 model was shown to have the capability of establishing distinguishable features from the 3D MR-ART data. Our baseline experiment with this model yielded a training accuracy of up to 95% after 100 epochs with the global dataset and showed that the model was definitively classifying data without resorting to guessing the same class every time. However, after running the federated averaging algorithm with this dataset, we saw no improvement in performance for the server. In fact, the server consistently reported a very poor validation accuracy (guessing the same class almost every time), as shown below in Figure 1. The six local centers did seem to consistently improve with training during the process as shown in Figure 2 below. This behavior is very unexpected compared to the behavior of federated learning in several natural image classification tasks, such as the classification of the MNIST dataset. Typically, in such datasets, not only do local models improve over averaging and iterations, but the centralized server average model seems to improve as well. Our hypothesis is that the variability of the dataset among each local center is greater in medical images than that of MNIST data. This hypothesis is supported by the slow learning rate seen in the 3D MR-ART image dataset compared to that in the MNIST dataset as shown in Figure 6 and Figure 7.

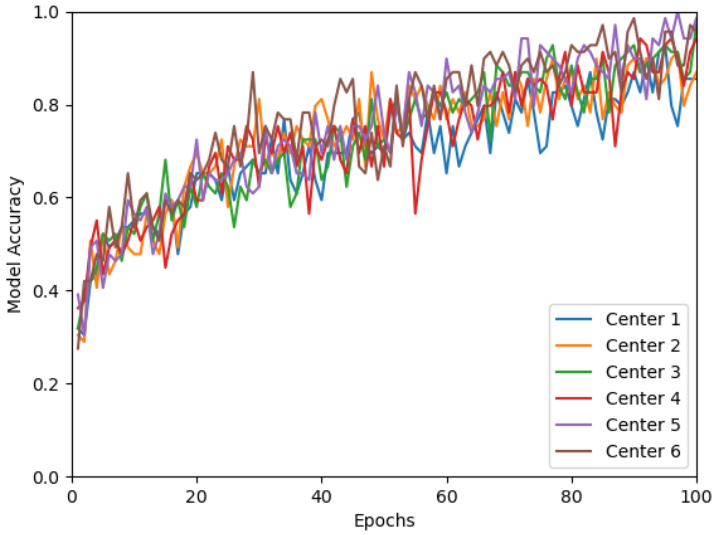

Figure 1: Center Training Accuracy of ResNet3D-18 on the 3D MR-ART Data with Federated Averaging

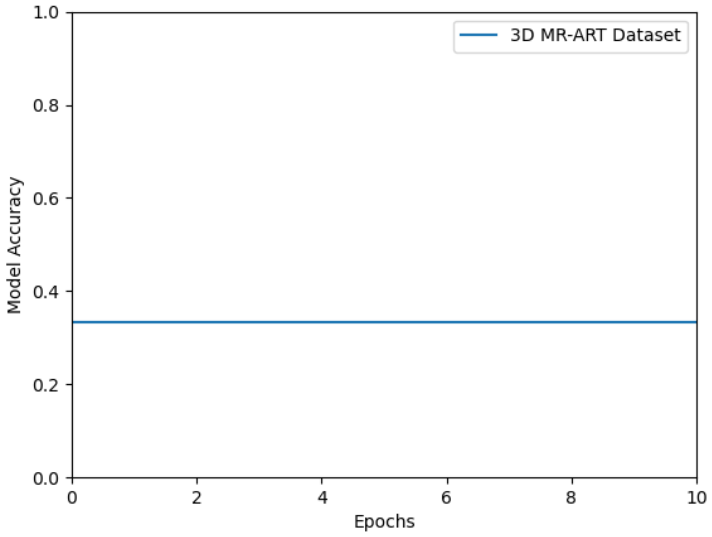

Figure 2: Server Validation Accuracy of ResNet3D-18 on 3D MR-ART Dataset after Federated Averaging

The 2D MR-ART dataset accomplished up to 95% training accuracy on the ResNet-18 model after 100 epochs. With these results confirming that the model can establish distinguishable features from the dataset, we proceeded with the federated averaging experiments. As found with the 3D MR-ART data and the ResNet3D-18 model, the federated averaging-generated server saw no improvement after 10 iterations of averaging, shown in Figure 4. Also seen in the previous experiment, the local center models seemed to consistently improve during the federated averaging process, with some reaching up to 94% in training accuracy, represented in Figure 3.

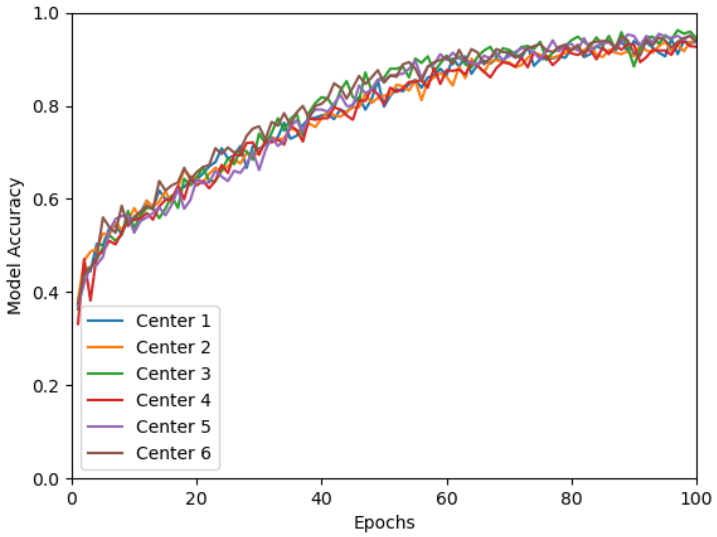

Figure 3: Center Training Accuracy of ResNet-18 on the 2D MR-ART Data with Federated Averaging

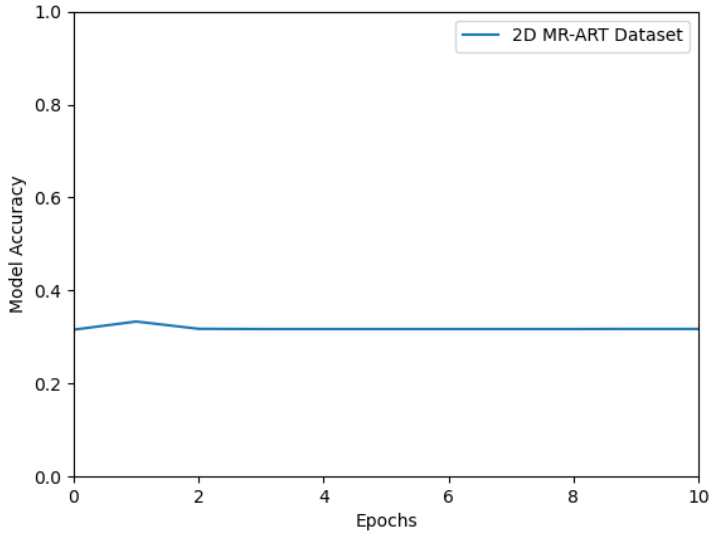

Figure 4: Server Validation Accuracy of ResNet-18 on 2D MR-ART Dataset after Federated Averaging

With our previous experiments resulting in poor federated averaging performance, we elected to shift towards testing less complicated models. We utilized the MNIST dataset and the basic Net2nn to establish the performance of the federated averaging algorithm. The server generated by federated averaging was shown to perform well on the MNIST data when trained using the Net2nn model. The server model reached up to 99% validation accuracy (Figure 7), and the centers showed an accuracy of up to 100% on their local datasets (Figure 5). When federated averaging was performed using the center-separated 2D MR-ART data and the Net2nn model, however, the Net2nn model proved to not be sophisticated enough to produce accurate classifications for the 2D MR-ART data. Results were poor for both the server and center models, and both seemed to fail to identify any distinguishing features within the data, as they each resorted to guessing one class for each input, as shown in Figures 6 and 7.

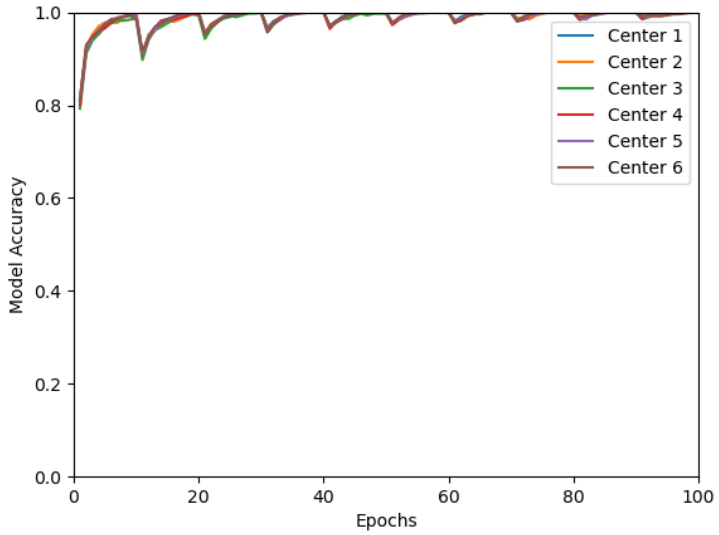

Figure 5: Center Training Accuracy of Net2nn on the MNIST Data with Federated Averaging

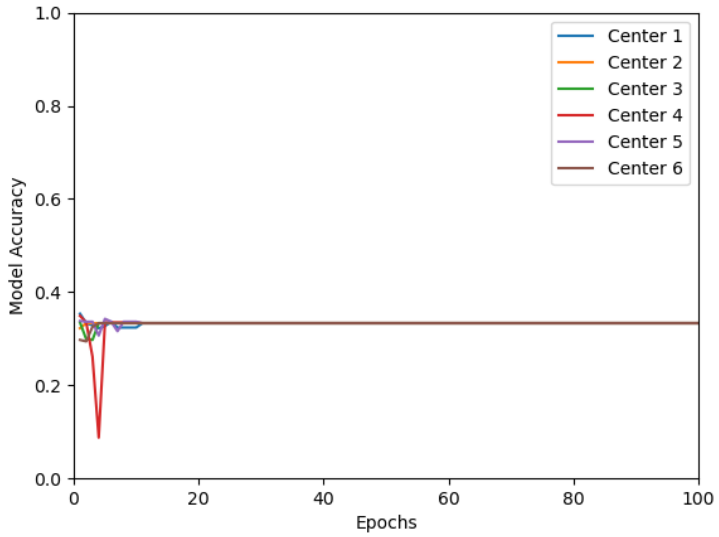

Figure 6: Center Training Accuracy of Net2nn on the 2D MR-ART Data with Federated Averaging

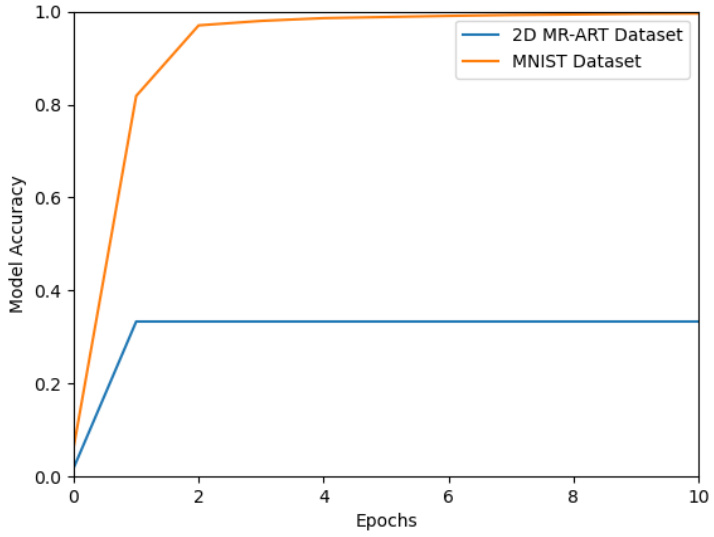

Figure 7: Server Validation Accuracy of Net2nn on MNIST and 2D MR-ART Dataset after Federated Averaging

Our success in producing high training and validation accuracy for both the centers and server using the Net2nn on the MNIST data and the failure of the Net2nn model to make any progress in classifying the 2D MR-ART images encouraged us to implement a slightly more complex model. We repeated the previous round of federated averaging experiments utilizing the same two datasets but using the adapted LeNet model instead of Net2nn. For the MNIST data, we found that while each of the individual centers produced high training accuracies for their local datasets (Figure 8), the federated averaging-generated server did not perform as well as it did when using the Net2nn model (Figure 10). While the server did seem to improve after the first two averaging iterations, the progress halted at about 23% validation accuracy, even while the servers continued to improve. The 2D MR-ART data produced variable results for center performance, as the training accuracy seemed to improve for all centers but one. For a few of the centers, accuracy reached up to 95%, represented in Figure 9. The server model, on the other hand, did not ever seem to make any improvements (Figure 10).

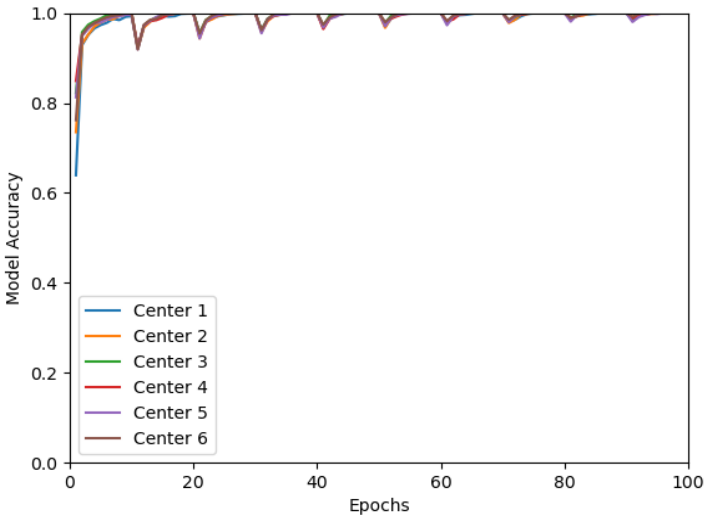

Figure 8: Center Training Accuracy of Adapted LeNet on the MNIST Data with Federated Averaging

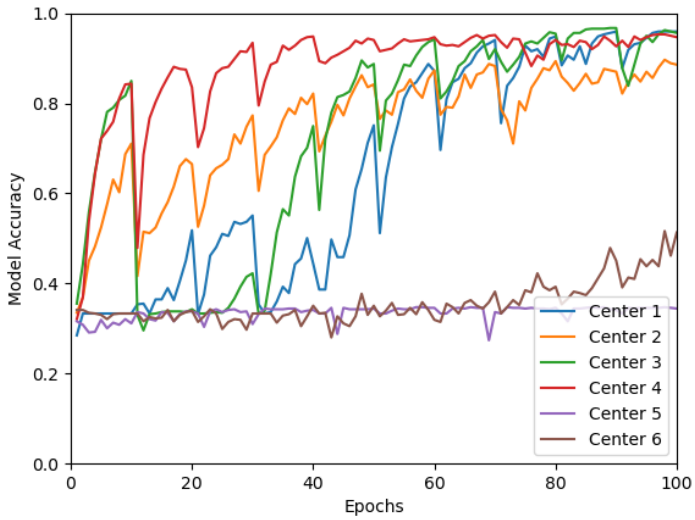

Figure 9: Center Training Accuracy of Adapted LeNet on the 2D MR-ART Data with Federated Averaging

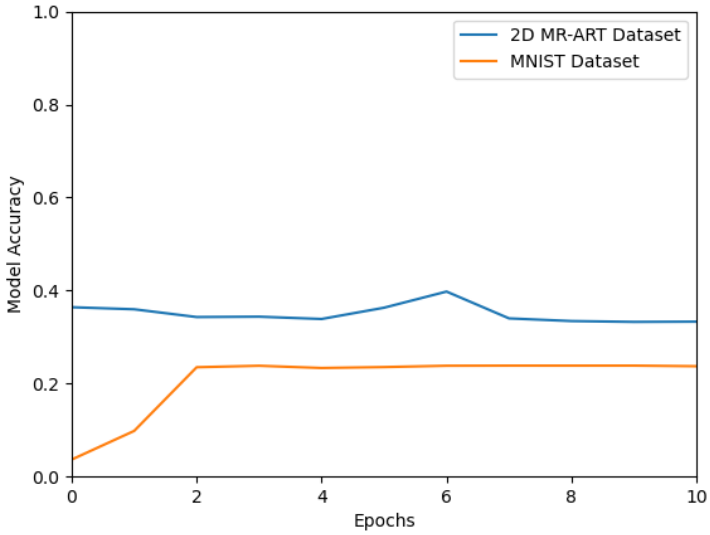

Figure 10: Server Validation Accuracy of Adapted LeNet on MNIST and 2D MR-ART Dataset after Federated Averaging

We took the saved models for each center from the federated averaging experiment and gathered the validation accuracy score for each center on each local dataset, shown in Table 1. We also utilized the same selection of local datasets to train the adapted LeNet model on each center's data for 100 epochs without federated averaging applied. These results are shown in Figure 11. We gathered the validation accuracy score for every center's saved model generated by this experiment for each local dataset as well. The scores of each center on all datasets are represented in Table 2.

	Validation Accuracy					
Centers \ Dataset	Center 1 Dataset	Center 2 Dataset	Center 3 Dataset	Center 4 Dataset	Center 5 Dataset	Center 6 Dataset
Center 1	0.95	0.8	0.79	0.78	0.8	0.77
Center 2	0.84	0.88	0.82	0.83	0.85	0.84
Center 3	0.74	0.79	0.94	0.73	0.75	0.75
Center 4	0.81	0.79	0.8	0.95	0.82	0.82
Center 5	0.34	0.34	0.34	0.34	0.34	0.33
Center 6	0.52	0.5	0.55	0.51	0.48	0.53

Table I. Validation Accuracy Scores of Each Center Trained on the Adapted LeNet with Federated Averaging for Every Local 2D MR-ART Dataset

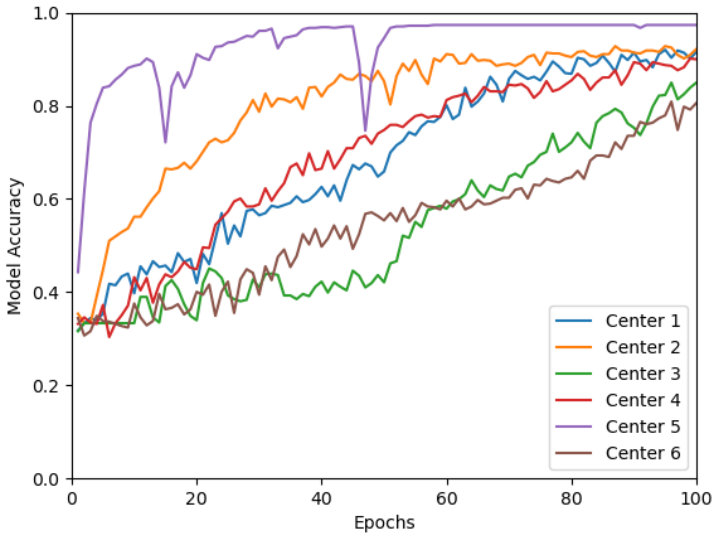

Figure 11:Center Training Accuracy of Adapted LeNet on the 2D MR-ART Data Without Federated Averaging

	Validation Accuracy					
Centers \ Dataset	Center 1 Dataset	Center 2 Dataset	Center 3 Dataset	Center 4 Dataset	Center 5 Dataset	Center 6 Dataset
Center 1	0.91	0.87	0.87	0.86	0.87	0.88
Center 2	0.86	0.9	0.83	0.84	0.84	0.85
Center 3	0.8	0.79	0.85	0.78	0.77	0.79
Center 4	0.78	0.78	0.79	0.83	0.78	0.8
Center 5	0.73	0.73	0.78	0.74	0.97	0.74
Center 6	0.77	0.77	0.79	0.75	0.75	0.84

Table II. Validation Accuracy Scores of Each Center Trained on the Adapted LeNet Without Federated Averaging for Every Local 2D MR-ART Dataset

VII. Discussion

We did not achieve our anticipated outcomes for the federated averaging-generated server model for any of the medical imaging datasets. In fact, we found only improvements when the MNIST data was used in training and averaging the parameters of a very simple model (Net2nn). This lends support to the implication that federated averaging may be a useful tool in generalizing the performance of models engaged in very simple classification tasks, as well as providing a layer of privacy protection to the local center's data. However, the same model that produced the desired results after federated averaging was applied using MNIST data was incapable of performing the complex computations necessary to classify our 2D MR-ART. When we implemented a more complicated model with more parameters (the adapted LeNet model), we found that while almost every individual center was seen to improve in accuracy across all training sessions, the federated averaging algorithm failed to produce a working model for both the MNIST and the 2D MR-ART data. The failure of the algorithm to produce an averaged server model that performed at any level of accuracy in these experiments implies that when too

many parameters are present in a model, the process of simple federated averaging is not enough to create a useful server model. There seems to be a fine line between implementing center models that are both sophisticated enough to perform the necessary classification tasks on complicated medical imaging datasets and center models that are simple enough to be useful to the federated averaging algorithm.

It was also interesting to note that during the process of federated averaging for the 2D MR-ART dataset using the adapted LeNet model, some of the centers achieved relatively high training accuracy, while a couple of the centers struggled to improve. One, notably, did not ever improve beyond the threshold of single-class guessing. The same datasets were used to train the adapted LeNet model without averaging, and each center reached high training accuracy scores on both their data and the other centers' datasets. This implies that in the process of federated averaging, if the model being trained has too many parameters for the algorithm to reliably use, the process of updating the center models with the averaged parameters has the potential to render them unable to ever learn distinguishing characteristics for their local data, or at the very least hindering their performance.

VIII. Conclusions and Future Work

Our goal of evaluating the performance of federated averaging on medical imaging data resulted in overall poor performance by the federated averaging-generated server model. The algorithm seemed to struggle to average the weights and biases of models with many parameters in a meaningful way, and we saw success for federated averaging only when using a model too simplistic to perform classification tasks with the medical imaging data in any capacity. The results of our experiments lead us to conclude that while the concept of federated learning would

have many benefits for use in the medical field, in theory, the simple federated averaging approach will need to be improved significantly in order to reliably use medical data, particularly in classification tasks.

References

[1] 27+ Most Popular Computer Vision Applications in 2022. (n.d.). Www.v7labs.com. https://www.v7labs.com/blog/computer-vision-applications

[2] Bass, E. (2022, January 4). The Unique Problems of Medical Computer Vision. Sirona Medical. https://sironamedical.com/the-unique-problems-of-medical-computer-vision/

[3] Brendan, M. H., Moore, E., Ramage, D., Hampson, S., & Blaise. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1602.05629

[4] Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

[5] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.

[6] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. Frontiers in Artificial Intelligence, 3. https://doi.org/10.3389/frai.2020.00004

[7] Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer Vision in Healthcare Applications. Journal of Healthcare Engineering, 2018, 1–4. https://doi.org/10.1155/2018/5157020

[8] Hara, K., Kataoka, H., & Yutaka Satoh. (2017). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1711.09577

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1512.03385

[10] Kataoka, H., Tenga Wakamiya, Hara, K., & Yutaka Satoh. (2020). Would Mega-scale Datasets Further Enhance Spatiotemporal 3D CNNs? ArXiv (Cornell University).

[11] McMahan, B., & Ramage, D. (2017, April 6). Federated Learning: Collaborative Machine Learning without Centralized Training Data. Google AI Blog. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[12] Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Medical Ethics, 22(1). https://doi.org/10.1186/s12910-021-00687-3

[13] Nárai, Á., Hermann, P., Auer, T. et al. Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans. Sci Data 9, 630 (2022). https://doi.org/10.1038/s41597-022-01694-8

[14] Patel, H. (2021, December 30). Data-Centric Approach vs Model-Centric Approach in Machine Learning. Neptune.ai. https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning

[15] Polat, E. I. (2020, September 28). Federated Learning: A Simple Implementation of FedAvg (Federated Averaging) with PyTorch. Medium. https://towardsdatascience.com/federated-learning-a-simple-implementation-of-fedavg-federated-averaging-with-pytorch-90187c9c9577

[16] Shukla, P. (2022, December 11). Machine Learning with Limited Data. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/12/machine-learning-with-limited-data/

[17] Singh, R. (2022, October 4). Recent Advances in Modern Computer Vision. Medium. https://towardsdatascience.com/recent-advances-in-modern-computer-vision-56801edab980

[18] Srivastava, A. (n.d.). Council Post: The Evolution Of Computer Vision And Its Impact On Real-World Applications. Forbes. Retrieved April 24, 2023, from https://www.forbes.com/sites/forbestechcouncil/2021/10/14/the-evolution-of-computer-vision-and-its-impact-on-real-world-applications/?sh=48564b48c6ab

[19] Tanaka, S.C., Yamashita, A., Yahata, N. et al. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci Data 8, 227 (2021). https://doi.org/10.1038/s41597-021-01004-8

[20] Teng, Q., Liu, Z., Song, Y., Han, K., & Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems, 28(6), 2335–2355. https://doi.org/10.1007/s00530-022-00960-4

[21] This is what makes deep learning so powerful. (2022, March 27). VentureBeat. https://venturebeat.com/datadecisionmakers/this-is-what-makes-deep-learning-so-powerful/

[22] Top 7 Computer Vision Use Cases in Healthcare in 2023. (n.d.). Research.aimultiple.com. Retrieved April 24, 2023, from https://research.aimultiple.com/computer-vision-healthcare/

[23] Voulodimos, A., Doulamis, N., Bebis, G., & Stathaki, T. (2018). Recent Developments in Deep Learning for Engineering Applications. Computational Intelligence and Neuroscience, 2018, 1–2. https://doi.org/10.1155/2018/8141259

[24] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[25] Zhang, S., Li, Z., Zhou, H.-Y., Ma, J., & Yu, Y. (2023). Advancing 3D medical image analysis with variable dimension transform based supervised 3D pre-training. Neurocomputing, 529, 11–22. https://doi.org/10.1016/j.neucom.2023.01.012